IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Missing Value Imputation-A Review

Dipalika Das[1*], Maya Nayak[2], Subhendu Kumar Pani[3]

[1]Department of MCA, TACT, BBSR, Odisha India
[2,3]Dept. of Computer Science and Engineering, OEC, BBSR, Odisha, India

Corresponding Author: dipalika.das@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i4.548558 | Available online at: www.ijcseonline.org

Accepted: 15/Apr/2019, Published: 30/Apr/2019

Abstract- The problems of missing values in the field of data mining have become emerging areas of research in recent years. It has been a great challenge in research for quite a long time. The missing values may occur due to several reasons. The missing values in the data set can affect accuracy and performance of result when any algorithm is implemented on it. Presence of missing values leads to less efficiency and difficulty in extracting meaningful information. As we go through the literature we can find there are various imputation techniques basing on type of missing value. Since the amount of data is increasing day by day, there is a need for an appropriate technique to handle the missing values in the data set. In this paper a brief year wise study of existing methods are being done so that it would be a great help while formulating and implementing a new one towards solving the problem of missing values.

Key words- Data Mining, Data Set, Missing Values, Algorithm, Information, Imputation

I. INTRODUCTION

As we know, missing value is a common problem in the area of data mining research. It creates problem during analysis and processing of data leading to less efficiency. In order to overcome this problem researchers have adopted several strategies. As we move through the literature study we can find there are improvements in existing methods towards getting better streamlined and smooth running solving technique.

II. RELATED WORK

Qian Ma et al. (2018) proposed Order-Sensitive Imputation for Clustered Missing values (OSICM) framework, in which missing values were imputed sequentially such that the values filled earlier in the process were also used for later imputation of other missing values(MVs). They found the order of imputation and effectiveness and efficiency of OSICM framework were interdependent. They searched the optimal imputation order as an optimization problem. They formulated an algorithm (linear greedy algorithm) to find the exact optimal solution and proposed two approximate/heuristic algorithms. They also conducted extensive experiments on real and synthetic datasets to show the importance of OSICM framework.

Teresa Pamula et al. (2018) focused on gathering real traffic data from the road network. This introduced application of multilayer perceptrons and deep learning networks using auto encoders for the prediction task to be performed. Here the aim of experiment was to investigate the sensitivity of neural networks to loss of input data for predicting traffic flows in a road network and to propose a strategy for substituting lost data for preserving the accuracy of prediction.

Siamak et al. (2018) considered missing feature values in the context of optimal Bayesian classification which selected a classifier that minimized the expected error with respect to the posterior distribution governing an uncertainty class of feature-label distributions. This paper derived a closed form decision rule for the optimal Bayesian classifier and utilized Hamiltonian Monte Carlo for the Gaussian model with arbitrary covariance matrices. Better performance was seen when compared to linear discriminant analysis, quadratic discriminant analysis, and support vector machines in conjunction with Gibbs sampling imputation using synthetic and real-world omics data. Hamiltonian Monte Carlo optimal Bayesian classifier (OBC-HMC) was evaluated basing on microarray and RNA-seq breast cancer data sets.
Wang et al. (2018) proposed a regularized local learning based method for missing value imputation. Here L_2 Regularized Local Least Squares imputation model (RLLSimpute_L_2) was imposed between the target gene and its neighbors for estimating the missing values of the target genes. L_1 Regularized Local Least Squares imputation model (RLLSimpute_L_1) was being introduced. Towards end, experiments were conducted on six microarray datasets and RLLSimpute_L_2 and RLLSimpute_L_1 were compared

with nine state-of-the-art imputation methods in terms of three evaluation metrics. Here experimental results showed the effectiveness of regularization techniques in removing the risk of over-fitting and the superiority of RLLSimpute_L_2 over its competing ones in missing value imputation.

Wujun Si et al.(2018) proposed a reliability model for repairable systems with incomplete failure time data under general repair. The proposed model was suitable for modeling both a single system and multiple systems subject to failure heterogeneity. A maximum likelihood estimation method was developed to estimate the model parameters. Based on the proposed model, statistical inference of the missing times was developed. Simulation and real-life case studies were conducted to verify the developed methods. In the proposed model two renewal distributions were considered for the modeling of system namely Gamma and IG. One MLE method was developed for model parameter estimation. Statistical inferences were drawn basing on proposed model.

Nur Afiquah Zakaria et al. (2018) experimented on missing data contained in air quality measurement data. They had used 4 imputation methods like Mean Top Bottom, Linear Regression, Multiple Imputation and Nearest Neighbour. Performance parameters like Mean Absolute Error, Root Mean Squared Error, Coefficient of Determination and Index of Agreement were used to find the appropriate imputation method to be used. Here Mean Top Bottom Imputation method was selected for filling in the missing values in air pollutants data. Linear Regression was found to be the worst imputation method. Nearest Neighbour method performed better than Multiple Imputation methods but less efficient as compared to Mean Top Bottom Imputation method.

Xu et al.(2018) proposed a new missing value imputation algorithm based on evidence chain (MAIEC) which could mine all relevant evidence of missing values in each data tuple and then combined this relevant evidence to build the evidence chain for further estimation of missing values. For large scale data processing they applied the Map-Reduce programming model to realize the distribution and parallelization of MAIEC. Experimental results showed that their proposed approach had higher imputation accuracy compared with the missing data imputation algorithm based on naïve bayes, mode imputation algorithm and the proposed missing data imputation algorithm based on K-nearest neighbour(KNN). MAIEC had higher imputation accuracy and was still better with the increase in missing rate. MAIEC was also proved to be suitable for the distributed computing platform and could achieve an ideal speedup ratio.

Zeng Yu et al. (2017) proposed a locally auto-weighted least squares imputation method for estimating missing values, which could automatically weigh the neighboring genes based on the importance of the genes in regression analysis. A new step was added to the basic LLSimpute method to update the neighboring genes weights based on the previous regression. A new weight calculation formula was used that minimized the square of Frobenius norm of residual matrix. An accelerating strategy was added to the LAW-LSimpute method in order to improve the convergence. An iterative missing value estimation framework of LAW-LSimpute (ILAW-LSimpute) was designed in this paper. Comparison was made between the performance of ILAW-LSimpute and the other five popular imputation methods. As per the result the new algorithm was able enough to reduce the estimation error. Here the experiment was on 8 real data sets.

Ivan Markovsky (2017) proposed an approach for data-driven filtering and control that combined the identification and the model-based design into one joint problem. The signal of interest was modeled as a missing part of a path of the data generating system. Ultimately, the missing data estimation problem was reformulated as a mosaic-Hankel structured matrix low-rank approximation / completion problem. A local optimization method, based on the variable projections principle, was then used for its numerical solution. The missing data estimation approach and the solution method as proposed here were illustrated on filtering and smoothing examples.

Weiwei Shi et al.(2017) proposed novel methods to improve the performance of missing data prediction in multivariable time series data. They illustrated how the smoothness constraints could be designed and how the correlation information in a sensor network could contribute to the missing data prediction in multivariable time series. They explained how to design matrix factorization objective functions. Finally they had implemented and verified the proposed methods with three real world data sets and one synthetic data set. For big data analysis they had used and tested the proposed methods on Apache Spark platform.

R. Misir et al.(2017) worked on a rarely used Hungarian dataset for heart disease from UCI repository. They chose LOCF,Mean-Mode substitution, IV- for missingness and multiple imputation technique for target dataset. The result obtained from imputed dataset was approximately 96%-100%. They suggested multiple imputation boot strapping techniques as a solution to work in the uncertain environment regarding missing value attributes in the input platform.

Malay Mitra et al.(2017) compared 3 imputation methods: Arithmetic Mean, Regression, Multiple Imputation using EMB algorithm. Their comparison based on 3 datasets from UCI repository containing MAR (Missing At Random) values. Here imputation accuracy was measured by Root Mean Square Error (RMSE). They found that there was no universal imputation method performing best in every situation.

Yelipe UshaRani et al. (2016) proposed handling missing values in medical datasets. They also put light on dimensionality reduction of medical data set using simple approach. While performing dimensionality reduction care was taken not to miss any attribute information. The proposed approach was applicable for categorical and numerical attributes. Normalization technique could be used if required. Imputation approach could be extended for prediction and classification of unknown medical records for predicting disease levels or symptoms through soft computing techniques.

Darryl et al.(2016) proposed a missing value imputation framework using stratified machine learning techniques. The proposed approach they used was FURIA (Fuzzy Unordered Rules Induction Algorithm) with stratification as a missing values imputation for real life incomplete cardiovascular datasets. Here the results were compared with non-stratified machine learning based missing value imputation methods and statistical mean-mode imputation methods using decision tree, SVM, K-NN. The experimental results proved that the proposed stratified machine learning methods were better. The proposed method was suitable for a big dataset having a large number of attributes with missing values.

Tejal(2016) proposed Naive Bayes Classifier for classification. He used the K-nearest neighbour algorithm (KNN) to handle with Missing values in Software Engineering datasets and found most of the missing value techniques were used to serve software development effort estimation techniques.

Y.Usha et al.(2016) proposed a novel imputation framework for missing values imputation. Their approach of filling missing values was rooted on class based clustering approach and aimed at dimensionality reduction of medical records. This dimensionality reduction method was used for carrying prediction and classification analysis. This was implemented on a case study to show how imputation was performed using proposed method. The proposed approach was applicable to impute both categorical and numerical attribute values. This could also be used for disease prediction.

Naveen et al. (2016) proposed a method to handle the uncertainty of missing values in the real world data sets. The research work removed inconsistency of the missing data sets by using fuzzy based K-NN. Here 3 different datasets (Hepatitis, Breast Cancer, Lung Cancer) were used. It was found that by introducing different levels of missing ratio in three different data sets the proposed method performed better.

Chi et al.(2016) proposed a K-POD method which was a simple extension of K-means clustering for missing data which worked even when the missingness mechanism was unknown, external information was unavailable and when there was significant missingness in data. K-POD method was found to be a simple, reliable and fast alternative to deletion and imputation in clustering missing data.

Jain et al.(2016) proposed the proper handling of missing data values and their analysis with removal of the anomalous data. The work focused on imputing missing values for numerical attribute in time series data set. This method provided more accurate and efficient result and reduced biasness of result for filling missing data. Theoretical analysis and experimental results showed that proposed method was more reliable as compared to other mean imputation technique for data analysis in the data mining field.

Lucia et al. (2015) used a medical database from Hospital de Santa Maria which was a small database with a reduced number of patients that contained biomarkers short time series with uneven sampling times and a high number of missing values. This work proposed three novel approaches based on fuzzy clustering to deal with incomplete short time series data, with even and uneven sampling times. The results were subsequently used for survival analysis, in order to find relations between biomarkers and patient-related times. These methods were implemented and validated using benchmark and artificially generated datasets. The proposed approaches in this work showed relevant results and proved efficient in dealing with datasets where data was composed by offsets in terms of amplitude. The tests performed on time series containing missing values provided significant results when compared to similar methods in literature. It was found through survival analysis of the biomarker clusters that several correlations existed between medical status of patients and patient-related times. The methods in the work could be applied in oncology studies, where biomarkers were extensively used to evaluate a patient's condition, in order to find useful knowledge not explicitly present in the data. The proposed approach could be extended to other medical related studies like gene expression data.

Saravanan et al. (2015) proposed a hybrid method that used a reliable machine learning technique known as support vector regression and a genetic algorithm was used with fuzzy possibilistic clustering to estimate missing values. Entire data were clustered based on their similarity and both fuzzy and possibilistic principles were used during clustering. Here each missing value became a member of more than one cluster centroids, which yielded more sensible imputation results. This paper used two datasets with different characteristics and the cluster size and the weighting factor parameters were optimized according to the corresponding dataset. As observed better imputation accuracy was achieved by FPCM-SVRGA (Fuzzy-Possibilistic c means using Genetic Algorithm with Support Vector Regression) as compared to the FCM-SVRGA method. The performance of experimental result demonstrated that the Fuzzy Possibilistic C-Means SVRGA imputation yielded a more sufficient, sensible estimation accuracy ratio for suitable clustering data.

Elsiddig et al.(2015) focused on missing data treatment on cluster performed on Sudan's Household Health Survey. Initially missing data mechanism and treatment rules were presented using the multiple imputation procedures. Two Step Cluster Analysis was chosen over a wide range of approaches of statistical pattern-recognition available for clustering. There was a risk of over-fitting of the data which must be considered because cluster analysis was a multivariable statistical technique. Like multi-variable statistical techniques any observation with missing data was excluded in the Cluster Analysis. Hence before performing the cluster analysis, missing values would be imputed using multiple imputations. The clustering results would be displayed in tables. The descriptive statistics and cluster frequencies would be produced for the final cluster model, while the information criterion table would display results for a range of cluster solutions.

Huseyin et al.(2015) introduced a framework for complete treatment of localized data corruptions due to severe noise sources. They proposed a novel algorithm to detect and localize possible corruptions from a given suspicious data and a posteriori estimator to impute the corrupted data. They also proposed a novel distance measure and shown to be superior in separating the corruptions. Their algorithm first splited the suspicious instance into parts through a binary partitioning tree in the space of data attributes and iteratively tested those parts to detect local anomalies using the nominal statistics extracted from an uncorrupted (clean) reference data set. Here the proposed framework was tested over several well-known machine learning data sets with synthetically generated corruptions and experimentally shown to produce remarkable improvements in terms of classification purposes with strong corruption separation capabilities. The proposed algorithm outperformed the typical approaches and were robust to varying training phase conditions.

Edgar et al. (2014) compared four methods to treat missing values in supervised classification problems. They chose the case deletion technique (CD), the mean imputation (MI), the median imputation (MDI) and the k-nearest neighbor (KNN) imputation. Here two classifiers were used: the Linear Discriminant Analysis (LDA) (parametric) and the KNN classifier (non parametric). There was no much difference between the results obtained with mean and median imputation. There was some difference between MI/MDI and KNN imputation only when a KNN classifier was used. However there was a noticeable difference between case deletion and all the imputation methods considered. CD performed badly, mostly due to the distribution of the missing values in a high percentage of instances. Overall KNN imputation seemed to perform better than the other methods because it was most robust to bias when the percentage of missing values increase.

Artur et al.(2014) presented the comparison of several clustering algorithms for data sets with missing values. They analysed the preprocessing techniques and specialized algorithms for data sets with missing values. The experiments showed that for moderate missing ratio (<10%) it was more advantageous to use pre-processing method (the best was median imputation). For high missing ratios (>25%) the specialized algorithms should be used, but it was worth mentioning that simple marginalization could elaborate better results than imputation in preprocessing. The results seemed to be independent of the type of data loss: missing from the whole data set, missing from only one cluster or attribute. The proposed dissimilarity measure seemed to be in concordance with the cluster quality indices.

Minakshi et al.(2014) used a student data set in this work. To impute the missing values they used three techniques that were: Lit Wise Deletion, Mean Imputation, and KNN Imputation. After applying these techniques they had 3 imputed data sets. They applied classification algorithms C4.5 or J48 on these imputed data sets. This work analysed the performance of imputation methods using C4.5 classifier. After that it was decided which imputation method would be the best to handle the missing value. On the basis of experimental results it was found that KNN had better accuracy than other two methods. Here Weka data mining tool was used for this analysis.

Xiaoping et al.(2014) proposed many methods to handle missing data that occurred frequently in longitudinal data analysis. Simulations under various situations were conducted to evaluate the performance of four most frequently used methods such as: Complete case (CC), mean substitution (MS), last observation carried forward (LOCF), and multiple imputation (MI). It was found from results that LOCF had the largest bias and the poorest 95% coverage probability in most cases under both MAR and MCAR missing mechanisms. So LOCF could not be used for longitudinal data analysis. CC and MI methods performed equally well under MCAR missing mechanism. MI had the smallest bias, smallest RMSE and best 95% coverage probability under MAR missing mechanism. Hence CC or MI method was the appropriate method to be used under MCAR while MI method was more reliable under MAR.

Kaiser(2014) presented simple methods for missing values imputation like using most common value, mean or median, closest fit approach and methods based on data mining algorithms like k-nearest neighbor, neural networks and association rules. He discussed their usability and presented issues with their applicability on examples. The missing values imputation method selection must be done by considering the structure of given dataset attributes. There were methods that suited more to numeric attributes and some suited symbolic attributes. These methods could also be combined.

Tapas et al.(2013) proposed the effect of missing values on data classification. They presented a comparative analysis of data classification accuracy in different scenarios. They had conducted an experiment on a dataset from UCI machine learning repository by using WEKA data mining tool to compare 5 most commonly used classification models on the effect of missing values classification performance. The classifiers showed quite satisfactory result when data instances with missing values were removed. When the missing values were filled by filters the classifiers showed unsatisfactory performance. They found that filling of missing values had no such effect on the performance of classifiers.

Luciano et al.(2013) analysed the influence of missing data on datasets when applied to traditional classification algorithms in data mining. They used ten UCI datasets and manipulated them to hold controlled levels of missing data. They found that classification performance decreased after significant insertion of missing values in all datasets. Analysis showed that Naive Bayes and SMO(Sequential Minimal Optimization) were least influenced by missing data where as IBK (Instance-based k-nearest Neighbors) was found to be most influenced, presenting the lowest accuracy, mostly in datasets whose independent variables were continuous.

Jing et al.(2013) proposed a missing data completion method known as CBGMI(Cluster Based Gray Relational Analysis using Multiple Imputation). They separated non missing value instances from missing ones and clustered them. Each missing value instance was merged into the closest cluster through gray relational analysis based distance metric. They experimented on UCI and aerospace datasets and found that their algorithm was superior to other existing approaches.

Sujatha R.(2013) considered the database with missing values, identified the attribute type. If it was a continuous attribute, mean pre-imputation was applied otherwise (for discrete attribute) mode pre-imputation was applied. Then kernel function was applied separately to both the attributes. Mixture kernel function was obtained by integrating both the discrete and continuous kernel function. Now iterative kernel estimator was applied separately for continuous as well as discrete attributes to get final value for imputation. This data was imputed in the missing dataset to make it as a complete dataset. Data pre-processing occurred using clustered algorithm. Performance analysis was done by comparing imputed values without using clustering and using clustering.

Aasha M.(2013) used datasets that were subjected to pre-processing techniques and the missing values were imputed using K-NN, Frequency estimator method, RBF(Radial Basis Function) kernel and Polynomial kernel and then a mixed kernel (RBF kernel and Poly kernel), and a spherical kernel mixed with poly kernel and a spherical kernel mixed with RBF kernel. Finally the performance of these imputation methods were evaluated using Root Mean Square Error(RMSE) and Correlation Coefficient. The result showed that the proposed approach was better than the existing imputation methods in terms of co relation coefficient and root mean square error (RMSE).

Santosh et al.(2013) proposed to impute the missing values in a mixed attribute data set. They used kernel functions for the discrete as well as continuous attributes and then a mixture kernel function was proposed by combining a discrete kernel function with a continuous one. Further an estimator was constructed based on the mixture kernel. They used a higher order kernel for missing value estimation was spherical and polynomial kernel. These kernels provided maximum functionalities on higher dimensional space. In another method mixture kernel function i.e. spherical kernel function was in linear combination with RBF kernel. These mixture kernel functions provided better extrapolation and interpolation. They had chosen completed data sets from UCI repository for experiment. They had generated some missing value in the percent of 10,20,30,40 for each data set. The proposed method was evaluated with some traditional algorithm like Frequency estimator (FE), polynomial kernel, and RBF kernel. The result showed that the proposed approach was better than those existing imputation methods in terms of classification accuracy and root mean square error (RMSE) at different missing ratios.

Ji Liu et al.(2013) proposed an algorithm to estimate missing values in tensors of visual data. The algorithm worked even with a small amount of samples and it could propagate structure to fill larger missing regions. Here methodology was built on recent studies about matrix completion using the matrix trace norm. First, they proposed a definition for the tensor trace norm that generalized the established definition of the matrix trace norm. Second, similar to matrix completion, the tensor completion was formulated as a convex optimization problem. The straightforward problem extension was significantly harder to solve than the matrix case because of the dependency among multiple constraints. To tackle the problem, they developed three algorithms:

Simple Low Rank Tensor Completion (SiLRTC), Fast Low Rank Tensor Completion (FaLRTC), and High Accuracy Low Rank Tensor Completion (HaLRTC). The SiLRTC algorithm was simple to implement and employed a relaxation technique to separate the dependant relationships and used the Block Coordinate Descent (BCD) method to achieve a globally optimal solution; the FaLRTC algorithm utilized a smoothing scheme to transform the original non-smooth problem into a smooth one and could be used to solve a general tensor trace norm minimization problem; the HaLRTC algorithm applied the Alternating Direction Method Of Multipliers (ADMMs) to the problem. Their experiments showed potential applications of their algorithms and the quantitative evaluation indicated that their methods were more accurate and robust than heuristic approaches. The efficiency comparison indicated that FaLTRC and HaLRTC were more efficient than SiLRTC and between FaLRTC and HaLRTC the former was more efficient to obtain a low accuracy solution and the later was preferred if a high-accuracy solution was desired.

Bhavisha et al.(2012) had discussed all the imputation methods for finding the missing values from the dataset. They had also presented the comparative review as well as advantage and disadvantage of the different imputation methods for finding the missing value from the dataset in the field of data mining.

R.Devi Priya et al.(2012) proposed a new technique based on Genetic Algorithm and Bayes' Theorem to impute missing discrete attributes which often occured in real world applications. The experimental results clearly showed that the proposed approach significantly improved the accuracy rate of imputation of the missing values. It worked better for datasets even with missing rates as high as 50% when compared with other existing methods.

Noel et al.(2012) proposed a Neural Selective Input Model (NSIM) that accommodated different transparent and bound models, while providing support for NNs to handle MVs directly. By embedding the mechanisms to support MVs they could obtain better models that reflected the uncertainty caused by unknown values. Their experiments on several UCI datasets with both different distributions and proportion of MVs showed that the NSIM approach was very robust and yielded good to excellent results. They demonstrated the usefulness and validity of the NSIM, making this a first-class method for dealing with the problem. Their future work would exploit the possibility of using selective inputs on other types of NNs and would try to extend this work to radial basis functions and recurrent networks.

J. Luengo et al. focused on fuzzy rule based classification systems (FRBCSs) considering 14 different approaches to missing attribute values treatment that are presented and analyzed. The analysis involved three different methods, in which they distinguished between Mamdani and TSK models. From the obtained results, the convenience of using imputation methods for FRBCSs with missing values was stated. The analysis suggested that each type behaved differently while the use of determined missing values imputation methods could improve the accuracy obtained for these methods. Thus, the use of particular imputation methods conditioned to the type of FRBCSs was required.

Satish et al.(2012) proposed a missing value imputation method based on K-Means and nearest neighbors. This method was applied on clinical datasets from UCI Machine Learning Repository. This method used the imputed objects for further imputations. The results showed that proposed method performed better than simple method (without using imputed values for further imputations) but it was not the case for all the datasets as error in earlier imputation might propagate to further imputations. There was scope for several new missing value imputation methods based on using imputed values for later imputations.

K. Raja et al.(2012) proposed the analysis of broadly used methods to treat missing values which were either continuous or discrete. And then, an estimator was advocated to impute both continuous and discrete missing target values. The proposed method was evaluated to demonstrate that the approach was better than existing methods in terms of classification accuracy. It was found that missing values filled with better accuracy leads to better results.

Ganga A. R et al.(2012) proposed a mixture kernel-based iterative non-parametric estimator based on higher order kernel functions for data sets having both continuous and discrete independent attributes was designed. It utilized all available observed information, including observed information in incomplete instances (with missing values), to impute missing values, whereas existing imputation methods used only the observed information in complete instances (without missing values). In future, this work could be extended to check with different combinations of kernel methods to achieve better results. Also the kernel based approach could be extended to find missing values for heterogeneous attribute data sets.

Ibrahim et al. (2012) proposed a new approach which utilized machine learning and artificial intelligence systems. They experimented by using four different datasets. Their work dealt with the estimation of missing data through novel techniques. The estimation system involved an auto-associative model to predict the input data, coupled with the k-nearest neighbors to approximate the missing data. The results showed that the hybrid NN-KNN could be applied when the record had more than one missing value in a row. On the contrary, the KNN algorithm was implemented for the same problem. The results showed that NN-KNN method was able to produce better imputation accuracy. The findings also showed that the method seemed to perform better in cases where there was dependency among the variables.

Somasundaram et al. (2012) proposed a new method RMS(Refined Mean Substitution) imputation method which was being implemented and evaluated on WDBC data set. The performance of the missing value imputation algorithms was measured with respect to different percentage of missing values in the data set. The performance of reconstruction was compared with the original WDBC data set. Here the proposed algorithms provided better performance than the most popular and standard method. The performance of the algorithms was evaluated with five different metrics. In almost all the cases their proposed algorithms performed better than MC(Most Common Attribute Value)/mean value substitution method.

III. TABULAR COMPARISON OF PREVIOUS 5 YEARS ALGORITHMS

Table 1

Sl. No	Year Of Publication	Author Name	Algorithm	Dataset	Performance
1	2018	Teresa Pamuła	Neural Network	Traffic Dataset	MLPs(Multi layer Perceptron) give better prediction results than DLNs (Deep Learning Network)
2	2018	Siamak Zamani Dadaneh, Edward R. Dougherty	Hamiltonian Monte Carlo optimal Bayesian classifier (OBC-HMC)	Breast Cancer Dataset	OBC-HMC outperforms other classifiers(Support vector machine (SVM),Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA))
3	2018	Qian Ma, Yu Gu, Wang-Chien Lee, Ge Yu,	Order-Sensitive Imputation for Clustered Missing values (OSICM) framework	Air Quality and Wiki4HE Datasets.	OSICM has better classification accuracy and scalability than existing methods
4	2018	Nur Afiqah ZAKARIA, Norazian Mohamed NOOR	Mean Top Bottom Method	5 Air pollutant Datasets, 3 Meteorological Datasets	It gave smallest error for all percentages of missing data
5	2018	Aiguo Wang, Ye Chen, Ning An, Jing Yang, Lian Li,and Lili Jiang	Regularized Local Least Squares imputation model(RLLSimpute_L2)	6 microarray gene datasets (3time series 3 non time series datasets)	It outperforms its competitors by achieving smaller imputation errors and better structure preservation of differentially expressed genes.
6	2018	Xiaolong Xu, Weizhi Chong, Shancang Li, Abdullahi Arabo	Missing data Imputation approach Based on a Chain of Evidence (MIBOCE)	UCI Dataset, American Census Dataset	The imputation accuracy of MIBOCE is better than that of other algorithms.
7	2018	Wujun Si, Qingyu Yang , Leslie Monplaisir, and Yong Chen	Maximum Likelihood Estimation Method(MLE)	The reliability analysis of an electronic equipment/ system in a manufacturing plant consisting of multiple identical	It performs effectively in terms of the parameter estimation accuracy, and the performance is robust with respect to the missing ratio.

				electronic modules/components	
8	2017	Zeng Yu, Tianrui Li, Shi-Jinn Horng, Yi Pan, Hongjun Wang, and Yunge Jing	Locally Auto-Weighted Least Squares Imputation (Law-Lsimpute)Method An iterative missing value estimation framework of LAW-LSimpute (ILAW-LSimpute)	8 datasets (SP.Alpha, SP.Cdc, Infection, Ronen, Y.Calcineurin, Lymphoma, Gasch, Ogawa)	As compared with the performance of the other five popular imputation methods, it is found that the new algorithm is able to reduce the estimation error.

9	2017	Weiwei Shi, Yongxin Zhu, Philip S. Yu, Jiawei Zhang, Tian Huang, Chang Wang, and Yufeng Chen	5 models:
			MFS: Matrix Factorization with Smoothness constraints;
			CSM: Correlated Sensors based Matrix factorization;
			USM:Uncorrelated Sensors based Matrix factorization;
			CSMS: Correlated Sensors based Matrix factorization with Smoothness constraints;
			USMS:Uncorrelated Sensors based Matrix factorization with Smoothness constraints;
				Three real-world data sets(Motes,Sea-Surface Temperature,Gas Sensor Dataset) and one synthetic data set.	The proposed five models show much better performance than the baseline methods

10	2017	Malay Mitra, R.K.Samanta	EMB Algorithm(EM Algorithm with boot starp approach	Breast Cancer, kidney,hepatitis Dataset	Bivariate dataset requires imputation with regression, multivariate dataset requires multiple imputation
11	2017	R. Misir, R.K. Samanta	Bootstrap Algorithm with MI,LOCF,Mean Mean-mode substitution and IV	Hungarian Data Set	Multiple imputation bootstrapping technique is a solution to work with missing value attributes.
12	2016	Yelipe UshaRani, Dr.P.Sammulal	Imputation approach based on clustering to fix missing values, dimensionality reduction and classification	Medical Dataset	This method is first of its kind which may be used to perform missing values imputation, classification, and disease prediction in a single stretch.
13	2016	Tejal Patil	Naive Bayes Classifier	SE Dataset	It is simple, performance is good, calculation efficiency is high
14	2016	Y.Usha Rani, P. Sammulal	Dimensionality reduction by using concept of mapping distance,	Medical Dataset	This method is capable to impute both categorical and numeric values of

			imputation through class based clustering approach		medical records.
15	2016	R. Naveen Kumar and M. Anand Kumar	Fuzzy based K-NN	Breast cancer , Lung cancer, Hepatitis Dataset	The proposed method outperforms the remaining algorithm.
16	2016	Jocelyn T. Chi, Eric C. Chi, and Richard G. Baraniuk	K-POD method	Wine Dataset	It produced more accurate and fast result than other approaches.
17	2016	Swati Jain & Mrs. Kalpana Jain	Proposed method for imputing missing values for numerical attribute	Time series Dataset (Hydroelectric Generation in India 1965-2013, Average Global Temperature 1880-2014, U.S. Motor Gasoline Consumption 1950-2014, World Wood Production 1961- 2011)	The proposed method performed significantly better than all other methods.
18	2015	L_ucia Maria Pina Moreira Pires da Cruz	Three novel approaches based on fuzzy clustering to deal with incomplete short time series data, with even and uneven sampling times.	Medical Database of cancer patients	It provided significant results when compared to similar methods
19	2015	P.Saravanan, P.Sailakshmi	Genetic Algorithm with Support Vector Regression (SVRGA), fuzzy possibilistic c means(FPCM) clustering method	Iris and marine db	It provides better imputation accuracy than FCM-SVRGA
20	2015	Elsiddig Elsadig Mohamed Koko, Amin Ibrahim Adam Mohamed	Multiple imputation method	Sudan Household survey	It focuses on missing data treatment on cluster performed on survey data
21	2015	Huseyin Ozkan, Ozgun Soner Pelvan, and Suleyman S. Kozat	A novel framework for corruption detection, localization and imputation	Training Dataset	Improvements achieved by the proposed framework in several classification tasks
22	2014	Minakshi, Dr. Rajan Vohra, Gimpy	Performance of imputation methods(lit wise deletion, mean imputation, KNN imputation) using Classification algorithm C4.5.WEKA data mining tool was used for analysis	Student Database	Accuracy of KNN is greater than other two techniques
23	2014	Edgar Acuna and Caroline Rodriguez	Compares four methods to treat missing values:	12 Datasets(Iris, Hepatitis,	KNN imputation seems to perform better than the

			case deletion technique (CD), the mean imputation(MI), the median imputation (MDI) and the k-nearest neighbor (KNN) Imputation using two classifiers: the Linear Discriminant Analysis (LDA) and the KNN classifier.	Sonar,heartc, Bupa, Ionosphere, Crx,Breastw, Diabetes, Vehicle, German, Segment)	other methods because it is most robust to bias when the percentage of missing values increases.
24	2014	Artur Matyja, Krzysztof Simiński	Algorithm for calculation of dissimilarity index	Iris,Glass,Telugu Dataset	For moderate missing ratio (<10%) preprocessing method and for high missing ratio (>25%) specialized algorithms should be used. The result is independent of type of data loss (missing from whole dataset, cluster, attribute)
25	2014	Xiaoping Zhu	Four imputation methods: 1) complete case (CC); 2) mean substitution (MS); 3) last observation carried forward (LOCF);and 4) multiple imputation (MI). Simulation Approach	MCAR and MAR Datasets	CC method was superior to the MS, LOCF, and MI methods under MCAR missing mechanism while MI method was superior to CC, MS, and MI methods under MAR

IV. CONCLUSION

Missing values have always been a problem in extracting meaningful information from data set. In order to get accurate result we need to have a complete data set. Hence different techniques were published till date and still work is going on to find an improved technique to handle the missing values. By going through all the published papers we could say all imputation techniques were not applicable to all types of data set. The implementation of these techniques were based on type of data set and structure of attribute type. From this literature we could have a brief idea of year wise development in the field of imputation of missing values in data set.

REFERENCES

[1] Qian Ma, Yu Gu, Wang-Chien Lee and Ge Yu, ”*Order-Sensitive Imputation for Clustered Missing Values”*, IEEE Transactions on Knowledge and Data Engineering, 1041-4347 ©2018.

[2] Teresa Pamuła, “*Impact of Data Loss for Prediction of Traffic Flow on an Urban Road Using Neural Networks”*, IEEE Transactions On Intelligent Transportation Systems 1524-9050 © 2018.

[3] Siamak Zamani Dadaneh , Edward R. Dougherty and Xiaoning Qian , “*Optimal Bayesian Classification With Missing Values”*, IEEE Transactions On Signal Processing, Vol. 66, No. 16, August 15, 2018.

[4] Aiguo Wang, Ye Chen, Ning An, Jing Yang, Lian Li, and Lili Jiang, “*Microarray Missing Value Imputation: A Regularized Local Learning Method*”, IEEE, 1545-5963 ©2018.

[5] Wujun Si, Qingyu Yang , Leslie Monplaisir and Yong Chen, “*Reliability Analysis of Repairable Systems With Incomplete Failure Time Data*”, IEEE , 0018-9529 © 2018.

[6] Nur Afiqah Zakaria, Norazian Mohamed Noor,” *Imputation Methods For Filling Missing Data In Urban Air Pollution Data For Malaysia*”, Urbanism. Arhitectură. Construcţii, Vol. 9 , No. 2 , 2018.

[7] Xiaolong Xu, Weizhi Chong, Shancang Li, Abdullahi Arabo, “*Missing Data Imputation Based On The Evidence Chain*”, IEEE Access, Vol. 6, 2169-3536, 2018.

[8] Zeng Yu, Tianrui Li, Shi-Jinn Horng, Yi Pan, Hongjun Wang and Yunge Jing, “*An Iterative Locally Auto-Weighted Least Squares Method for Microarray Missing Value Estimation*”, IEEE Transactions On Nanobioscience, Vol. 16, No. 1, January 2017.

[9] Ivan Markovsky,” *A Missing Data Approach to Data-Driven Filtering and Control*”, IEEE Transactions On Automatic Control, Vol. 62, No. 4, April 2017.

[10] Weiwei Shi, Yongxin Zhu, Philip S. Yu, Jiawei Zhang, Tian Huang, Chang Wang, and Yufeng Chen, “*Effective Prediction of Missing Data on Apache Spark over Multivariable Time Series*”,

IEEE Transactions on Big Data ,DOI 10.1109/TBDATA.2017.2719703.

[11] R. Misir and R.K. Samanta,"*A Study on performance of UCI Hungarian dataset using missing value management techniques"*, IJCSE, Volume-5, Issue-3, 2017.

[12] Malay Mitra and R. K. Samanta," *A Study on Missing Data Management*", IJCSE, Volume-5, Issue-2, E-ISSN: 2347-2693, 2017.

[13] Yelipe UshaRani, Dr.P.Sammulal, "*An Innovative Imputation and Classification Approach for Accurate Disease Prediction*", International Journal of Computer Science and Information Security (IJCSIS), Vol. 14 S1, February 2016.

[14] Darryl ND, Rahman MM, "*Missing Value Imputation Using Stratified Supervised Learning for Cardiovascular Data*", Global J Technol Optim 7:6 DOI: 10.4172/2229-8711. S1:113,2016.

[15] Tejal Patil, "*Systematic Mapping Study of Missing ValuesTechniques using Naive Bayes*", IRJET, e-ISSN: 2395 -0056, Volume: 03, Issue: 03 , Mar-2016.

[16] Y.Usha Rani1, P. Sammulal, "*A Novel Approach for Imputation of Missing Attribute Values for Efficient Mining of Medical Datasets – Class Based Cluster Approach*", Rev. Téc. Ing. Univ. Zulia. Vol. 39, No 2, 184 - 195, 2016.

[17] R. Naveen Kumar, M. Anand Kumar, "*Enhanced Fuzzy K-NN Approach for Handling Missing Values in Medical Data Mining*", Indian Journal of Science and Technology, Vol 9(S1), DOI: 10.17485/ijst/2016/v9iS1/94094 , December 2016.

[18] Jocelyn T. Chi, Eric C. Chi, and Richard G. Baraniuk, "*k-POD A Method for k-Means Clustering of Missing Data*", arXiv:1411.7013v3 [stat.CO] 27 Jan 2016.

[19] Swati Jain & Mrs. Kalpana Jain, "*Estimation of Missing Attribute Value in Time Series Database in Data Mining"*, Global Journals Inc. (USA), Volume 16, Issue 5, Version 1.0, Year 2016.

[20] P.Saravanan,P.Sailakshmi, "*Missing Value Imputation Using Fuzzy Possibilistic C Means Optimized With Support Vector Regression And Genetic Algorithm*", JATIT & LLS, Vol.72, No.1, 2015.

[21] Elsiddig Elsadig Mohamed Koko, Amin Ibrahim Adam Mohamed, "*Missing Data Treatment Method On Cluster Analysis*", International Journal of Advanced Statistics and Probability, Vol.3,No.2 ,191-209, 2015.

[22] Huseyin Ozkan, Ozgun Soner Pelvan, and Suleyman S. Kozat, "*Data Imputation Through the Identification of Local Anomalies*", IEEE Transactions On Neural Networks And Learning Systems, Vol. 26, NO. 10, October 2015.

[23] Edgar Acuna ,Caroline Rodriguez, "*The treatment of missing values and its effect in the classifier accuracy*", Research Gate, DOI: 10.1007/978-3-642-17103-1_60, 2015 .

[24] Artur Matyja, "*Comparison of Algorithms for Clustering Incomplete Data, Foundations Of Computing And Decision Sciences*", Vol.39, No.2, DOI: 10.2478/fcds-2014-0007, ISSN 0867-6356, 2014.

[25] Minakshi, Dr. Rajan Vohra, Gimpy, "*Missing Value Imputation in Multi Attribute Data Set*", IJCSIT, Vol. 5 (4) , 5315-5321, 2014,.

[26] Xiaoping Zhu, "*Comparison of Four Methods for Handling Missing Data in Longitudinal Data Analysis Through a Simulation Study"*, Open Journal of Statistics, 4, 933-944, 2014.

[27] Jiri Kaiser, "Dealing with Missing Values in Data, Journal Of Systems Integration", 2014/1.

[28] Tapas Ranjan Baitharu and Subhendu Kumar Pani, "Effect of Missing Values on Data Classification, JETEAS", 4(2): 311-316, (ISSN: 2141-7016), 2013.

[29] Luciano C. Blomberg, Duncan Dubugras A. Ruiz, "*Evaluating the Influence of Missing Data on Classification Algorithms in Data Mining Applications*",SBSI,2013.

[30] Jing Tian, Bing Yu, Dan Yu, and Shilong Ma, "*Clustering-Based Multiple Imputation via Gray Relational Analysis for Missing Data and Its Application to Aerospace Field*", The ScientificWorld Journal, Article ID 720392, 10 pages, 2013.

[31] Sujatha.R, "*Enhancing Iterative Non-Parametric Algorithm for Calculating Missing Values of Heterogeneous Datasets by Clustering"*, IJSR Publications, Volume 3, Issue 3, March 2013.

[32] Aasha.M, "*Imputation in Mixed Attribute Datasets using Higher Order Kernel Functions*", IJIET, Vol. 2 Issue 3, ISSN: 2319-1058, June 2013.

[33] Santosh Dane, Dr. R. C. Thool, "*Imputation Method for Missing Value Estimation of Mixed-Attribute Data Sets"*, IJARCSSE, Volume 3, Issue 5, ISSN: 2277 128X, May 2013.

[34] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye, "*Tensor Completion for Estimating Missing Values in Visual Data*", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 35, No. 1, January 2013.

[35] Bhavisha Suthar, Hemant Patel, Ankur Goswami,"*A Survey: Classification of Imputation Methods in Data Mining*",IJETAE, ISSN 2250-2459, Volume 2, Issue 1, January 2012.

[36] R.Devi Priya, S.Kuppuswami, "*A Genetic Algorithm Based Approach for Imputing Missing Discrete Attribute Values in Databases"*, WSEAS Transactions On Information Science And Applications, E-ISSN: 2224-3402, Volume 9, Issue 6, June 2012.

[37] Noel Lopes, Bernardete Ribeiro, "*Handling Missing Values Via A Neural Selective Input Model*" Neural Network World 4/12, 357-370, ICS AS CR 2012.

[38] Julian Luengo, Jose A. Saez, Francisco Herrera,"*Missing data imputation for fuzzy rule-based classification systems*", 16:863–881 DOI 10.1007/s00500-011-0774-4, 2012.

[39] Satish Gajawada, Durga Toshniwal, "Missing Value Imputation Method Based on Clustering and Nearest Neighbours", International Journal of Future Computer and Communication, Vol. 1, No. 2, August 2012.

[40] K. Raja , G. Tholkappia Arasu ,Chitra. S. Nair, "*Imputation Framework for Missing Values, International Journal of Computer Trends and Technology*", volume3,Issue2,2012.

[41] Ganga.A.R, B.Lakshmipathi, "*Higher Order Kernel Function Algorithm for Imputing Missing Values*",IJARCS, Volume 3, No. 3, ISSN No. 0976-5697, May-June 2012,.

[42] Ibrahim Berkan Aydilek and Ahmet Arslan, "*A Novel Hybrid Approach To Estimating Missing Values In Databases Using K-Nearest Neighbors And Neural Networks*", International Journal of Innovative Computing, Information and Control, ISSN 1349-4198,Volume 8, Number 7(A), pp. 4705-4717, July 2012,.

[43] R.S. Somasundaram, R. Nedunchezhian, "*Missing Value Imputation using Refined Mean Substitution*", IJCSI, Vol. 9, Issue 4, No 3, July 2012.